Anatomy of scientific misconduct. Bibliometric analysis of the *Deja Vu* database.

Antonio García-Romero[1], José M. Estrada[2] and José M. Rodriguez-Vallejo[3]

[1] *agr33@salud.madrid.org*
UPIB, Agencia Laín Entralgo. Comunidad de Madrid. c/Gran Vía, 27, Madrid (Spain)
Department of Economics. Universidad Carlos III de Madrid. c/Madrid, 126, Getafe (Spain)

[2] *josemanuel.estrada@salud.madrid.org*
Health Sciences Library, Agencia Laín Entralgo. Comunidad de Madrid. c/Gran Vía, 27, Madrid (Spain)

[3] *josem.vallejo@hotmail.es*
Urology Service, Hospital Infanta Leonor (Madrid). c/Gran Vía del Este, 80, Madrid (Spain)

Abstract
Scientific misconduct is a worrying problem whose incidence is increasing due to the increasingly competitive research environment in many countries. *Deja Vu* is a publicly available database of highly similar citations identified by eTBLAST from PubMed. We performed a bibliometric analysis of 116 pairs of duplicated publications whose earlier papers were published in the period 2004-2005. We selected all the cases that have been confirmed by experts. Our aim was to determine to what extent the duplicates with shared authors (SA) are different from those with different authors (DA) in terms of citations and other relevant variables. To this extent, we hope to contribute to a better understanding of the scientific misconduct problem. Our results reveal that there is a clear differentiation between the two types of publications. In the case of papers with different authors, the duplicates received fewer citations than the duplicates with shared authors. Moreover, the DA duplicates are published with a delay of two years on average, one year more than that for SA duplicates. This pattern suggests that fraudulent scientists try to hide their scientific misconduct.

Introduction
Science and technology represent one of the main sources of wealth and health for mankind. For this reason many governments around the world consider R&D activities to offer the best solutions to the current economic crisis. It is well known that scientific research generates important returns both in economic and social dimensions. Within this context, health-related research plays a crucial role because it represents roughly 25% of total research expenditure in the world (Burke, 2008) and around 40% of global research output (Garcia Romero, 2008).
However, during the last decades, the costs associated with carrying out scientific research have risen significantly. For example, the number of scientists and their salaries has increased considerably in recent years (Austin, 2006). Moreover, the costs of scientific equipment have duplicated since 1990. Unfortunately, the increase in research budgets has not evolved at the same rate. These circumstances have provoked a scarcity of resources and a subsequent increase of competition within the scientific community. Although the ensuing competitive behavior has contributed to improve the quality of research proposals and their results, it has also had unintended consequences associated with the widening of the "publish or perish" culture and, perhaps, a rise in the incidence of scientific misconduct.
Any form of scientific misconduct can damage the public perception of science and the reputation of scientists. But this issue is not only of concern for scientists. As several surveys on the social perception of science and technology have shown, in developed countries the majority of citizens clearly support publicly-funded health-related research. This social relevance of research is also observed in the mass media –i.e.: The Economist, The Boston Globe – that not only covers successful stories of science, but also those about plagiarism or any other kind of scientific misconduct. News items of this type can cause significant damage to scientists' credibility.

The consequences of scientific misconduct go beyond the damage caused to scientists' reputations. Institutions where these scientists work can be negatively affected and, in the case of health research, the impact on patients could also be adverse. Moreover, scientific misconduct generates considerable financial costs, as has been demonstrated recently by Michalek et al. (2010). Their study concludes that the direct costs of a single case of scientific misconduct could reach US$ 525,000.

According to the Office of Research Integrity (ORI), an institution supported by the U.S. Public Health Service, research misconduct "means fabrication, falsification, or plagiarism in proposing, performing, or reviewing research, or in reporting research results" (ORI, 2010). Among other duties, the ORI has the responsibility of reviewing and monitoring research misconduct in the USA. To do so, the ORI has at its disposal several tools such as eTBLAST, a text similarity detection software which compares any text to a collection of other texts, particularly those found in Medline, and *Déjà vu*, a database of highly similar citations identified by eTBLAST (Errami et al., 2008a, 2008b).

Deja Vu comprises more than 70,000 pairs of similar citations. It is a widely recognized tool that has demonstrated its usefulness in the fight against scientific misconduct. For instance, thanks to this database many publications have been retracted by the journals where they were published. *Deja Vu* is a publicly available database whose entries are ordered pairs of similar publications, called "earlier" and "later" articles. Each entry is classified into different categories depending on its main caracteristics. An interesting attribute is that of authorship, according to which there are two types of duplicates: (i) publications with shared authors (SA) and publications with different authors (DA). Furthermore, entries are evaluated by a pair of experts that carry out the document categorization within a range of options. Moreover, the experts provide supporting information and comments that justify each judgment. Nevertheless, many of the duplicated documents identified by the software are, in fact, legitimate publications (i.e.: periodic reviews, periodic guidelines, specialized databases and specialized federal register citations). These types of publications are labeled as "SANCTIONED" by the experts. However, there are other types of duplicates that correspond to different types of scientific misconduct, such as multiple submission or self-plagiarism. As a consequence, several publications have been retracted by journals. The *Deja Vu* project is widely recognized and valued by scientific stakeholders. For instance, several papers based on data forthcoming from *Deja Vu* have been published in leading academic journals (Errami & Harold, 2008; Long et al., 2009; Dove, 2009).

The goal of the present research is to conduct a bibliometric analysis of publications included in the Deja Vu database. Among other questions, we wished to explore to what extent the duplicates with shared authors (SA) are different from those with different authors (DA) in terms of citations or impact factor. By performing this analysis we hope to contribute to a better understanding of the scientific misconduct problem.

Data and Methods

Data

The data used in this paper have been gathered from different sources. First, from *Deja Vu* we considered all the entries labelled as "EXAMINED". In order to facilitate the citation count, we selected all the entries whose earlier articles were published in the period 2004-2005, a total of 115 cases (i.e.: pairs of publications). Given this selection procedure, we have 80 cases with non-shared authors and 35 with shared authors. Second, the citation counts for each paper were gathered from the SCOPUS database. Third, from the SCImago Journal and Country Rank (SJR) we obtained the SJR of each journal and year.

Variables

For each pair of papers we considered the following variables: publication year; number of authors; SJR of the journal; citations on a by-year basis. Concerning data for this last indicator, we computed the aggregate number of citations using three different time windows: two, three and five years. In addition, *Deja Vu* also provides information regarding the time lag (in months) between each pair of papers, and two additional indicators of percentage of text coincidence between the earlier and later papers: similarity ratio (abstract) and full text similarity (whole paper).

Methods

Using SPSS statistical software, we performed a descriptive analysis of the main variables for each group of publications (with and without shared authors). Moreover, in order to determine if differences were significant, we used ANOVA or Mann-Whitney U tests depending on the distribution of each. Essentially, we used the former when the condition of variance homogeneity was satisfied, and the latter to test for significant differences between groups.

Results

Our preliminary results (Table 1) show a clear differentiation among the publications sharing authors (SA) with those that do not have any author in common (DA). We report in a separate manner the results for each type of duplicate.

Duplicates with shared authors (SA)

The later papers in the SA category are published after an average of 13.5 months, and they receive significantly less citations than the earlier ones. By countries, publications from the USA represent 29% of duplicated records. However, due to its huge scientific production, such a result cannot be associated with a higher level of scientific misconduct. Other countries well represented in the list are Thailand (12%), China, France and Japan (9%). The later papers with shared authors could correspond to a variety of situations, some of which cannot be considered as engaging in scientific misconduct. For instance, some of them are published in the same issue of a journal. However, within the SA category there are also cases of self-plagiarism or multiple submissions.

Duplicates with different authors (DA)

The later papers in the DA category are published 23 months on average after the earlier paper and receive significantly less citations than the earlier ones. The duplicates within this category have a higher degree of similarity both in abstracts (60.6% vs 52.4%) and particularly in full texts (76.3 vs 55.4%). This last result indicates that an average duplicate paper with different authors has 76.3% of its text in common with the original. By countries, the situation is quite different when we compare the country of earlier and later papers respectively. On the one hand, the countries that lead the earlier paper production are the USA (27%), Turkey (12%) and China (8%). On the other, the countries where more duplicates with DA are published are Iran (15%), China (13%) and Turkey (9%).

Table 1. Main indicators for papers with and without shared authors.

	Shared authors (SA) (n=35)		*Different authors (DA) (n=80)*	
	Earlier paper	Later paper	Earlier paper	Later paper
Authors	4.2	4.3	3.4	3.4
SJR	0.442	0.339	0.333	0.335

2-year citations	4.8	2.1	3.7	1.5
3-year citations	5.7	10.9	7.8	3.0
5-year citations	21.7	10.7	15.9	5.1
Time lag (months)	13.5		22.8	
Similarity ratio (%)	52.4		60.6	
Full text similarity (%)	55.4		76.3	

With regards to differences that were found to be statistically significant, Table 2 shows the p-values of ANOVA or Mann-Whitney U test scores.

Table 2. Significant differences between SA and DA papers.

Variable	Earlier paper	Later paper
Authors	**0.092***	**0.062***
SJR	0.398	0.970
2-year citations	0.398	0.242
3-year citations	0.647	**0.025****
5-year citations	0.486	**0.011****
Time lag (months)	**0.000*****	13.5
Similarity ratio	**0.004*****	52.4
Full text similarity	**0.007*****	55.4
*= p<0.1; ** = p <0.05; *** = p<0.01		

From these results it can be concluded that both SA and DA duplicates have significantly less citations than the earlier papers (Fig 1 and 2).

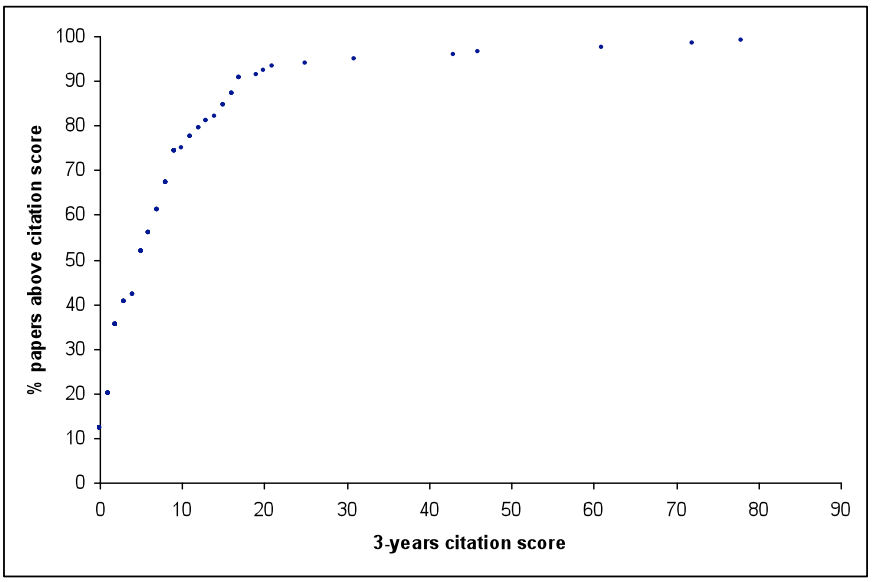

Figure 1. 3-year citation rates for earlier papers

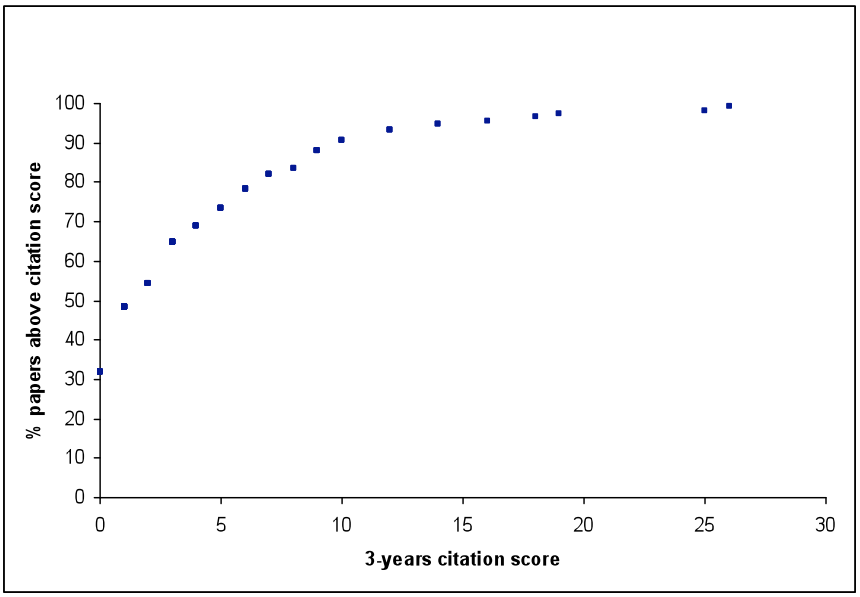

Figure 1. 3-year citation rates for later papers

Moreover, DA duplicates are published two years after the original, while SA duplicates are published one year on average later. Finally, the text similarity between original and duplicate publications is significantly higher for DA pairs.

Discussion and further research

These preliminary results suggest the existence of two different patterns citation and publication delay associated with the SA and DA categories, respectively. On the one hand, in the case of papers with different authors, the duplicates are published after two years on average and receive significantly fewer citations than the earlier paper that inspired them. On the other hand, those papers with shared authors seem to be more visible than papers with different authors and receive more citations. A possible explanation for this could be that some of these documents are legitimate publications.

Our preliminary results open interesting paths for further research that we intend exploring in the near future. Our first task will be to enlarge the dataset by including more entries from *Deja Vu*. We will also complete the journal impact indicators by taking into account their h-indexes. Moreover, a specific analysis of DA entries will be carried out in order to better understand the manner in which plagiarism occurs within the scientific community.

References

Austin, Jim (2006). Life scientists report rising salaries and high job satisfaction. Science, November 3.

Mary Anne Burke and Stephen A Matlin (eds.) (2008. Global Forum for Health Research, Monitoring Financial Flows for Health Research. Geneve.

SCImago. Journal and Country Rank. Retrieved March 11th, 2011 from: http://www.scimagojr.com/journalrank.php

Dove, Alan (2009). Regulators confront blind spots in research oversight. Nature Medicine, 15, 5, 469.

Errami, Mounir; and Garner, Harold (2008). A tale of two citations. Nature, 451, 7177, 397-399.

Errami, Mounir; Hicks, Justin M.; Fisher, Wayne; Trusty, David; Wren, Jonatahan D.; Long, Tara C.; and Garner, Harold R. (2008a). Déjà vu: A study of duplicate citations in Medline. Bioinformatics, 24(2), 243-249.

Errami, Mounir; Sun, Zhaohui; Long, Tara C.; George, Angela C; and Garner, Harold R. (2008b). Deja vu: a database of highly similar citations in the scientific literature. Nucleic Acids Research, 37, D921–4.

García Romero, A. (2008). Evaluación del impacto socioeconómico de la investigación biomédica: situación actual y perspectivas de futuro. Med. Clin (Barc). 131(Supl 5):1-5

Long, Tara C.; Errami, Mounir; Sun, Zhaohui; Garner, Harold R. (2009). Scientific integrity. Responding to possible plagiarism. Science, 323, 5919, 1293-1294.

Michalek, Arthur M.; Hutson, Alan D.; Wicher Camille P.; and Trump, Donald L. The costs and underappreciated consequences of research misconduct: A case study. PLoS Medicine, 7(9), e1000318.

ORI. Office of Research Integrity. Handling misconduct. Introduction [web page]. Retrieved December 1, 2010 from: http://ori.hhs.gov/misconduct/.